

Prédiction de revenus

A la recherche de prospects

Objectif

recherche de nouveaux clients

Qui ?

- Jeune
- Premier compte
- Futur hauts revenus

Où ?

Le monde entier

Comment ?

modèle prédictif

revenu parents

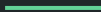
pays de l'individu

revenu moyen de son pays

indice de Gini de son pays



revenu de l'enfant



Sommaire

Données *pays*

- Préparation *mission 1*
- Analyse *mission 2*

Données *individus* *mission 3*

- Revenu des parents sachant celui des enfants ?
- Création des individus

Modèles statistiques *mission 4*

Sommaire

Données *pays*

- **Préparation** *mission 1*
- Analyse *mission 2*

Données *individus* *mission 3*

- Revenu des parents sachant celui des enfants ?
- Création des individus

Modèles statistiques *mission 4*

Données pays - Distribution des revenus

	country	year_survey	quantile	nb_quantiles	income	gdpppp
3300	FRA	2008	1	100	2958.3040	30357.0
3301	FRA	2008	2	100	4412.6753	30357.0

116 pays - *référence*

Centiles

→ classes de revenus

Avantages

- Inégalités intra-pays
- comparaison inter-pays

Valeurs manquantes

- GDP PPP pour 3 pays
- Quantile 41 de la Lituanie

Données pays - Indice de Gini

	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	...	2007	2008	2009
Country Code														
AGO	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	42.7	NaN
ALB	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	30.0	NaN

166 pays

5 pays manquants → **ajout manuel**

Gini de référence

- **Plus petit intervalle** d'années avec au moins **une valeur**
- **Moyenne** par pays

Données pays - Population

	1960	1961	1962	1963	1964	1965
Country Code						
ABW	5.421100e+04	5.543800e+04	5.622500e+04	5.669500e+04	5.703200e+04	5.736000e+04
AFG	8.996351e+06	9.166764e+06	9.345868e+06	9.533954e+06	9.731361e+06	9.938414e+06

263 pays & régions
du monde

Population de référence → **année 2011**

Population couverte ?

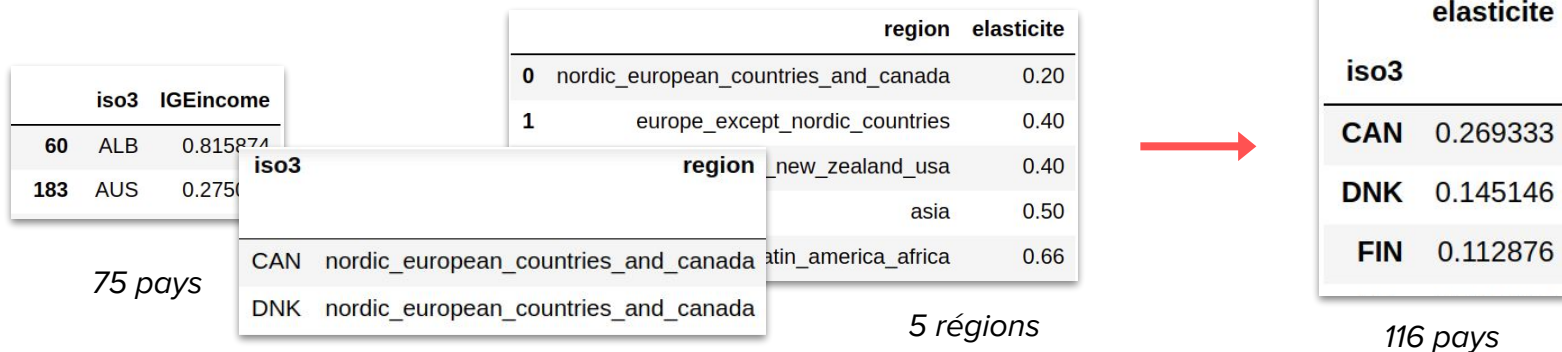
→ **jointure** gauche sur les **pays de référence**

→ **Gini & revenu** de **91%** de la **population**

Données pays - Coefficient d'élasticité

Élasticité intergénérationnelle

- mesure **l'impact** des revenu des **parents** sur celui de leurs **enfants**
- valeur de **0** (forte mobilité sociale) à **1** (faible mobilité sociale)



Sommaire

Données *pays*

- Préparation *mission 1*
- **Analyse** *mission 2*

Données *individus* *mission 3*

- Revenu des parents sachant celui des enfants ?
- Création des individus

Modèles statistiques *mission 4*

Analyse des pays - Sélection de pays types

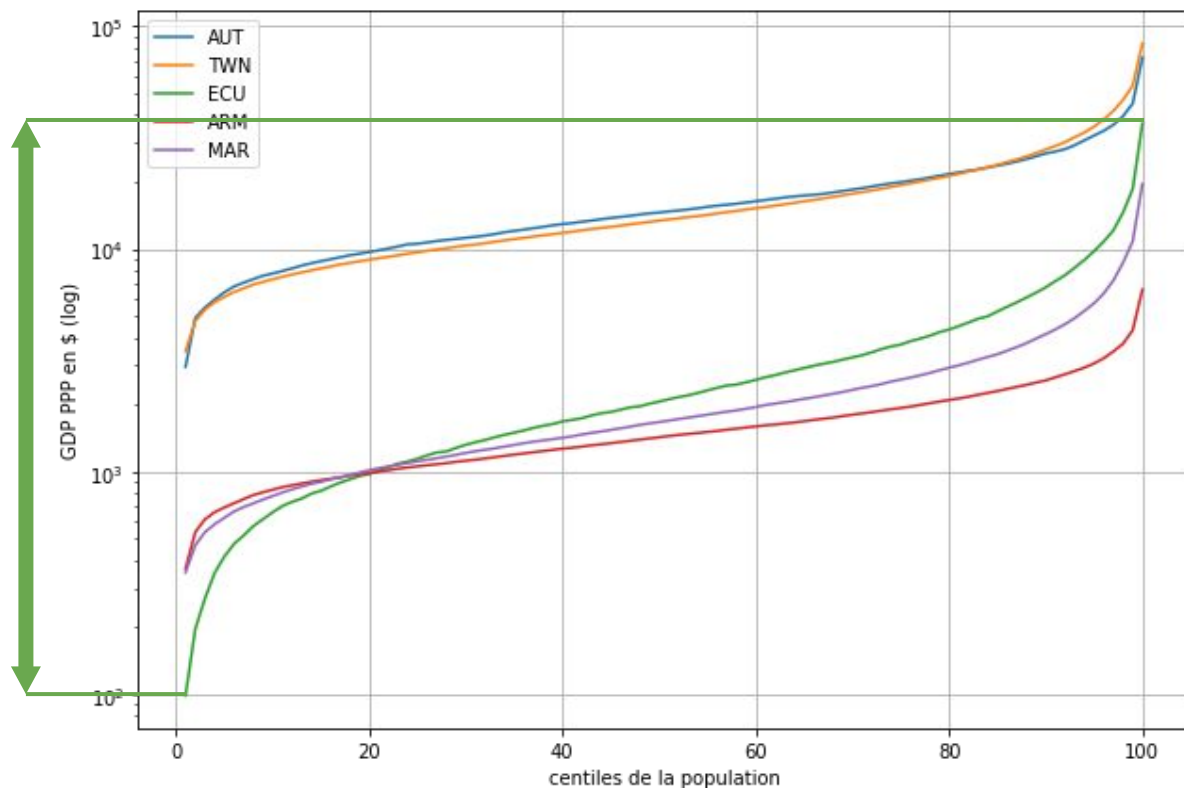
5 pays illustrant les inégalités ? → **Classification non-supervisée**

1. **dataframe** `code_pays` | *gini de référence* | *GDP PPP* → jointure sur `code_pays`
2. **normalisation** des données → `preprocessing.scale()`*
3. **kmeans** sur 5 clusters → `Kmeans()`*
4. on garde le pays le proche de chaque **centroïdes** → `pairwise_distance_agmin_min()`*

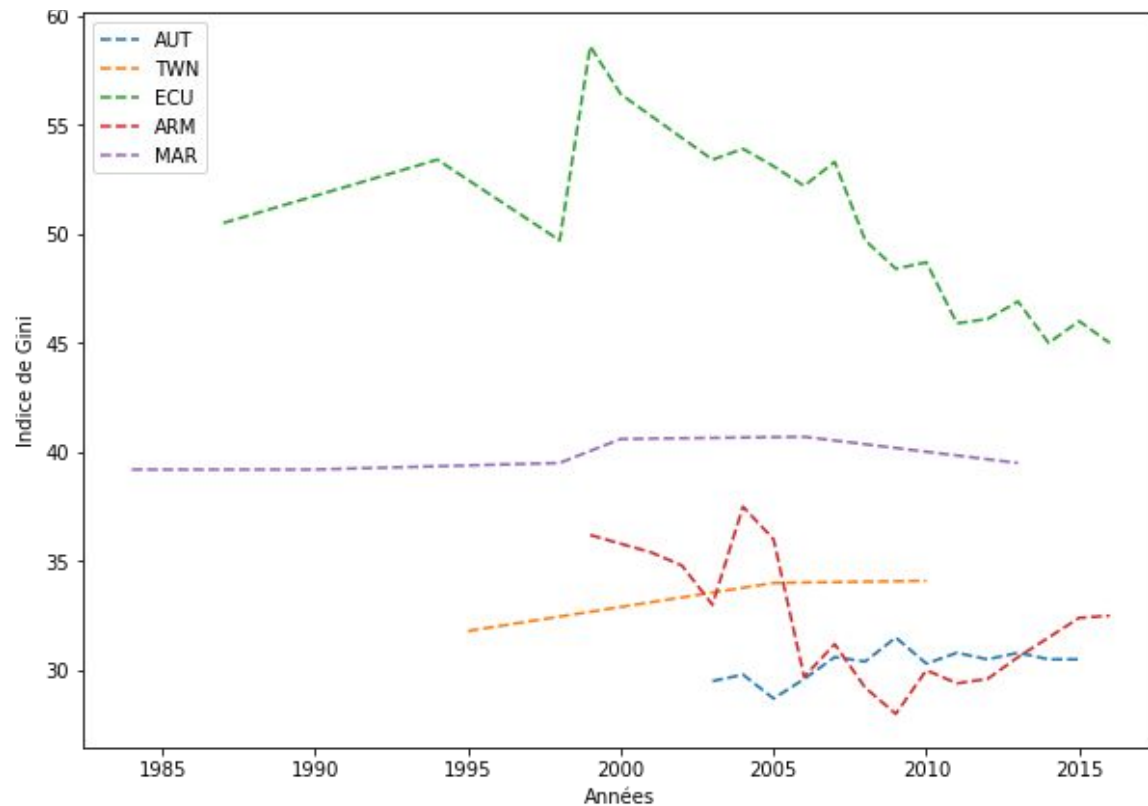
* librairie python sickitlearn

→ l'Autriche, Taiwan, l'Equateur, l'Arménie et le Maroc

Analyse des pays - Courbes de Lorenz



Analyse des pays - Indice de Gini



Analyse des pays - Indice de Gini

Plus fortes inégalités

Botswana - 62.0

Afrique du Sud - 61.7

Namibie - 61.1

Bélize - 57.7

Suriname - 57.6

Plus faible inégalités

Azerbaïdjan - 23.4

Slovénie - 25.3

Tchéquie - 26.4

Slovaquie - 26.7

Danemark - 26.7

La France

indice Gini de **32.1**

33ème pays dans le monde

Sommaire

Données *pays*

- Préparation *mission 1*
- Analyse *mission 2*

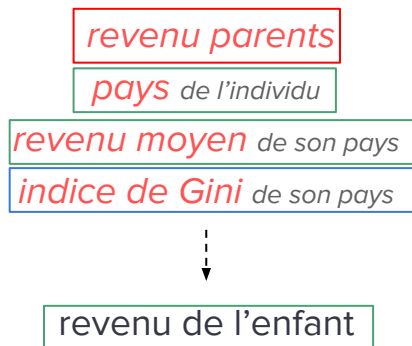
Données *individus* *mission 3*

- Revenu des parents sachant celui des enfants ?
- Création des individus

Modèles statistiques *mission 4*

Comment former les données d'individus ?

modèle prédictif



classe de revenu de l'enfant

	country	year_survey	quantile	nb_quantiles	income	gdpppp
3300	FRA	2008	1	100	2958.3040	30357.0
3301	FRA	2008	2	100	4412.6753	30357.0

→ 1 ligne = 1 individu

Comment déterminer la classe de revenu du parent ?

- depuis la classe de revenu enfant
- avec le coefficient d'élasticité

elasticite	
iso3	
CAN	0.269333
DNK	0.145146
FIN	0.112876

Sommaire

Données *pays*

- Préparation *mission 1*
- Analyse *mission 2*

Données *individus* *mission 3*

- **Revenu des parents sachant celui des enfants ?**
- Création des individus

Modèles statistiques *mission 4*

Revenu des **parents sachant** celui des **enfants**

→ **Fonction** permettant de calculer $P(c_{i,parent} | c_{i,child}, p_j = x)$

Formule liant le revenu d'un parent à son enfant

$$Y_{child} = e^{p_j \times \ln(y_{parent}) + \epsilon}$$

→ Génération de **couples de revenus parent/enfant**
pour **p_j** donné

Parents sachant enfants - Exemple

- on fixe p_j à **0.9**
- Génération gaussienne de **10 000** $\ln(Y_{parent})$ & erreur \mathcal{E}
- calcul de Y_{child} selon l'équation

```
y_child, y_parents = generate_incomes(n, pj)
```

- On fixe $nb_{quantiles}$ à **10** et on discrétise les revenus

```
sample = compute_quantiles(y_child, y_parents, nb_quantiles)
```

	y_child	y_parents	c_i_child	c_i_parent
0	0.783328	0.336317	5	2
1	1.617195	2.499539	7	9

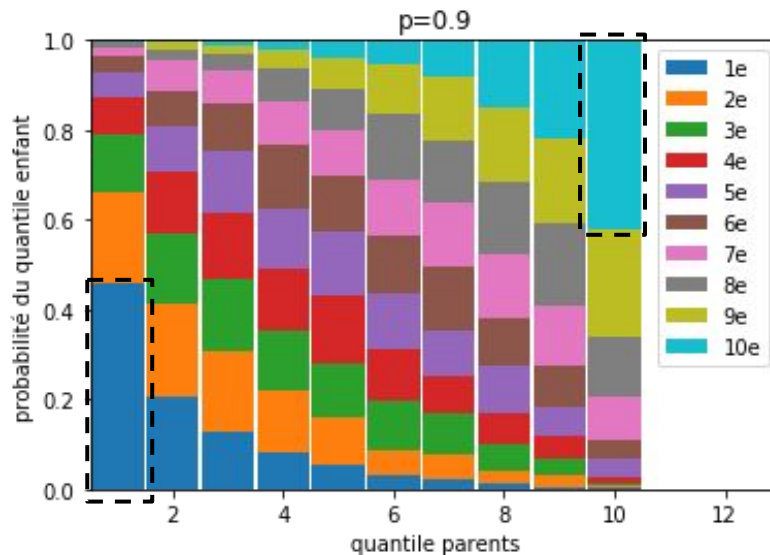
Parents sachant enfants - Exemple

c_i_child	c_i_parent
5	2
7	9

10 000 lignes

fréquence des $C_{i,parents}$ pour un $C_{i,child}$ fixé $\rightarrow P(c_{i,parent} | c_{i,child} = x, p_j = 0.9)$

```
cd = conditional_distributions(sample, nb_quantiles)
```



Parents sachant enfants - **Généralisation** de la démarche

```
def create_conditional_distributions_matrice(pj, nb_quantiles):  
    n = 1000*nb_quantiles  
    y_child, y_parents = generate_incomes(n, pj)  
    sample = compute_quantiles(y_child, y_parents, nb_quantiles)  
    return conditional_distributions(sample, nb_quantiles)
```

→ retourne une **matrice M** de dimension $\text{nb}_{\text{quantiles}} \times \text{nb}_{\text{quantiles}}$ où

$$M_{i,j} = P(c_{i,\text{parent}} = i | c_{i,\text{child}} = j, p_j = x)$$

Sommaire

Données *pays*

- Préparation *mission 1*
- Analyse *mission 2*

Données *individus* *mission 3*

- Revenu des parents sachant celui des enfants ?
- **Création des individus**

Modèles statistiques *mission 4*

Liste d'individus - Initialisation

	country	year	country	c_i_child	income_child	come	gdpppp
3300	FRA		0	ALB	1	728.89795	...
3301	FRA		1	ALB	2	916.66235	...
			2	ALB	3	1010.91600	...
			3	ALB	4	1086.90780	...

$$\begin{array}{ccccccc} 116 & \times & 100 & \times & 500 & = & 5\,800\,000 \\ \text{pays} & & \text{classes d'individus} & & \text{clones} & & \text{individus} \end{array}$$

Optimisation

- classes d'individus en *int8*
- Code pays en *Category*

177 → 60 Mo


Liste d'individus - Classe de revenu parents

`create_conditional_distributions_matrice()` → fréquence d'apparition des $C_{i,parent}$ sachant $C_{i,child}$

Comment créer **500** $c_{i,parents}$ depuis cette fréquence ?

```
def get_values_from_frequencies(n, values, frequencies):  
    nb_values = iteround.saferound(np.array(frequencies) * n, 0)  
    nb_values = [int(x) for x in nb_values]  
    return np.concatenate([ np.full(nb, value) for value, nb in zip(values, nb_values) ])
```

Problème : il faut retourner **n** valeurs

`get_values_from_frequencies(7, [a, b, c], [1/3, 1/3, 1/3])`  `[a, a, b, b, c, c]` **6 X**
`[a, a, a, b, b, c, c]` **7 OK**

Liste d'individus - Classe de revenu parents

Pour chaque **pays**:

p_j = coefficient d'élasticité du pays

```
distribution_conditionnelle = create_conditional_distributions_matrice(p_j, 100)
```

Optimisation → Calculé en **amont** & résultat **stocké** dans un fichier

Optimisation → *country* et *c_i_child* mis en **multi-index**

affectation des **c_i_parents** aux individus du **pays** et du **c_i_child** de l'itération

	country	c_i_child	income_child	c_i_parent
1222019	CZE	45	7011.11670	6
2386553	IRN	74	6818.36430	20

Liste d'individus - Finalisation

→ ajout de l'indice de Gini et GDP PPP

country	income_child	c_i_parent	gdpppp	gini_ref
HUN	5533.59230	79	18004.0	29.646154
KEN	720.09247	12	1429.0	46.580000
FRA	13469.08700	50	30357.0	32.092308

modèle prédictif

revenu parents

pays de l'individu

revenu moyen de son pays

indice de Gini de son pays



revenu de l'enfant

Sommaire

Données *pays*

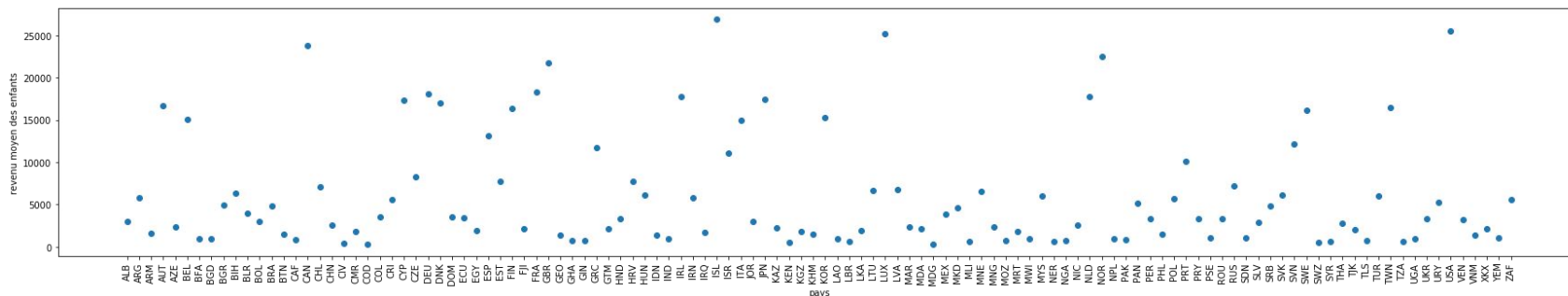
- Préparation *mission 1*
- Analyse *mission 2*

Données *individus* *mission 3*

- Revenu des parents sachant celui des enfants ?
- Création des individus

Modèles statistiques *mission 4*

Y a-t-il une **différence de revenu** entre les **pays** ?



ANOVA → income_child à partir de **pays**

ols() de statsmodel

f-valeur = 4935

p-valeur = 0.00

$\eta^2 = 0.50$

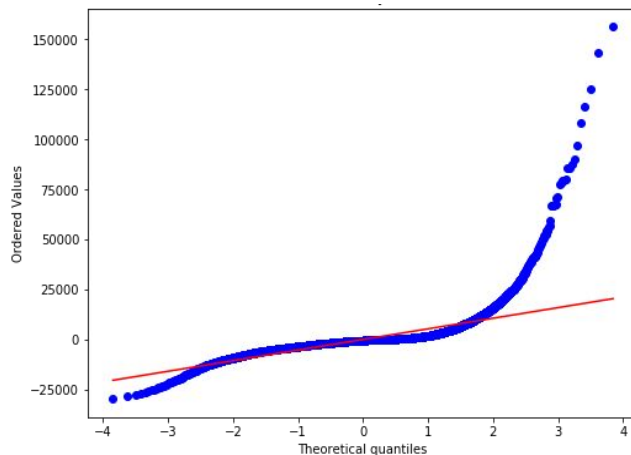
Il y a une **différence de revenu** entre les **pays**

Selon l'ANOVA, **50%** de la **variation de revenu**
est expliquée par le **pays d'origine**

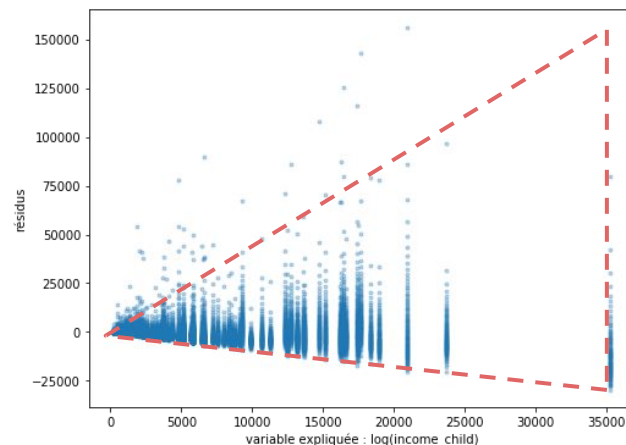
Le **pays** de naissance influe t'il sur le **revenu de l'individu** ?

Régression linéaire → **income_child** à partir de **GDP (PPP)**, **gini_ref**

`statsmodel.OLS()`



Les **résidus** ne semblent pas **gaussiens**

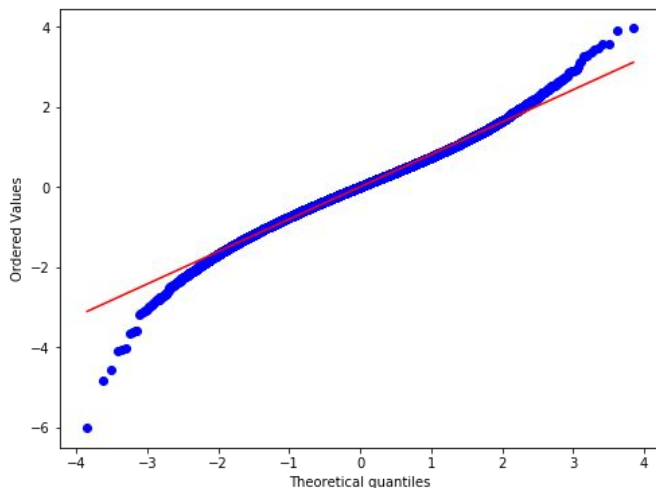


Les **résidus** n'ont pas de **variance constante**

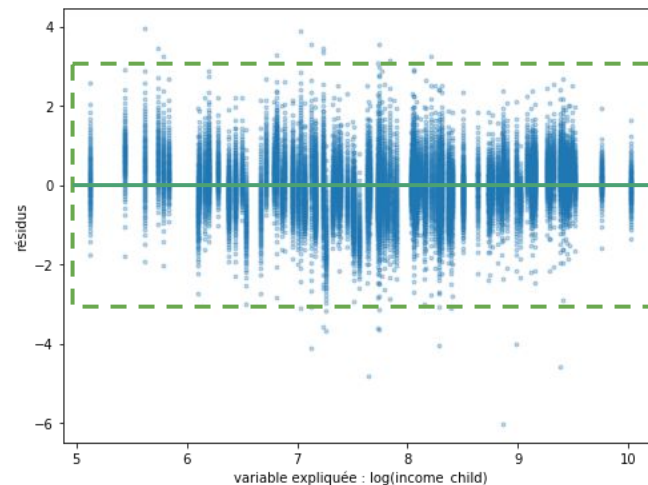


Le **pays** de naissance influe t'il sur le **revenu de l'individu** ?

Régression linéaire → **log(income_child)** à partir de **log(GDP PPP)**, **gini_ref**



Les **résidus** semblent **gaussiens**



Les **résidus** ont une **variance constante**

Le *pays* de naissance influe t'il sur le *revenu de l'individu* ?

Dep. Variable:	income_child	R-squared:	0.654			
Model:	OLS	Adj. R-squared:	0.654			
Method:	Least Squares	F-statistic:	5.473e+06			
Date:	Thu, 24 Jan 2019	Prob (F-statistic):	0.00			
Time:	13:53:02	Log-Likelihood:	-7.0274e+06			
No. Observations:	5800000	AIC:	1.405e+07			
Df Residuals:	5799997	BIC:	1.405e+07			
Df Model:	2					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	0.9657	0.003	278.182	0.000	0.959	0.972
gdpppp	0.8586	0.000	2973.284	0.000	0.858	0.859
gini_ref	-0.0177	4.34e-05	-407.754	0.000	-0.018	-0.018

Le **pays** de naissance influe t'il sur le **revenu de l'individu** ?

Régression linéaire → **log(income_child)** à partir de **log(GDP PPP)**, **gini_ref**

Facteurs d'inflation de la variance

variable	FIV
GDP PPP	1.12
gini_ref	1.12

Les variables sont
linéairement indépendantes

Distance de Cooks

`get_influence()` du modèle

$$seuil = \frac{4}{nb_{ind} - nb_{variable}} = 6.9e - 07$$

5.8% des individus sont influents

→ on les garde

Le **pays** de naissance influe t'il sur le **revenu de l'individu** ?

Régression linéaire → **log(income_child)** à partir de **log(GDP PPP)**, **gini_ref**

$$\text{SCT} = \text{SCE} + \text{SCR}$$

11 062 517 7 231 145 3 831 372

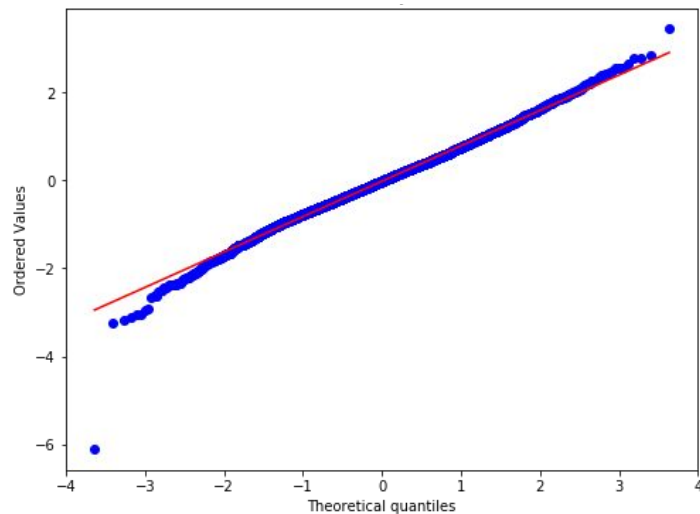
Selon le modèle,

$$R^2 = \text{SCE} / \text{SCT} = \mathbf{0.65}$$

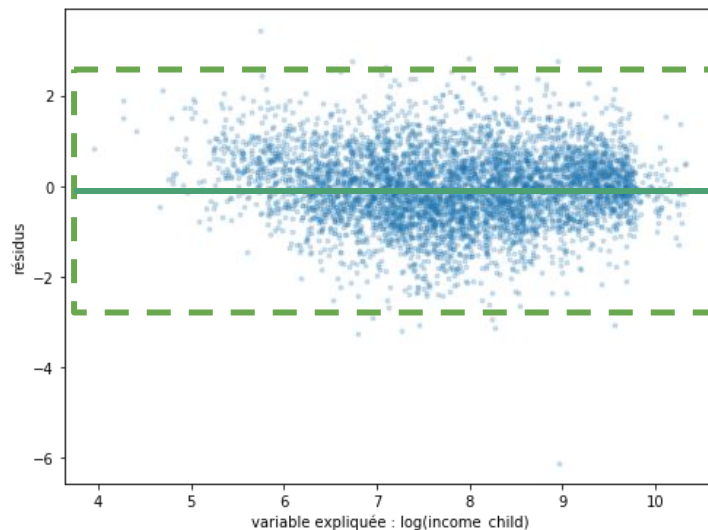
65% de la **variation de revenu** est expliquée par
le **salaire moyen** et l'indice de **Gini** du **pays de naissance**

Et si l'on ajoute la classe de **revenu des parents** ?

Régression linéaire → **log(income_child)** à partir de **log(GDP PPP)**, **gini_ref** et **log(C_{i,parent})**



Les résidus semblent gaussiens



Les résidus ont une variance constante

*Et si l'on ajoute la classe de **revenu des parents** ?*

Régression linéaire → **log(income_child)** à partir de **log(GDP PPP)**, **gini_ref** et **log(C_{i,parent})**

Facteurs d'inflation de la variance

variable	FIV
GDP PPP	1.12
gini_ref	1.12
C _{i,parent}	1.00

Les variables sont
linéairement indépendantes

Distance de Cooks

5.8% des individus sont influents

→ on les garde

Et si l'on ajoute la classe de **revenu des parents** ?

Dep. Variable:	income_child	R-squared:	0.700			
Model:	OLS	Adj. R-squared:	0.700			
Method:	Least Squares	F-statistic:	4.509e+06			
Date:	Thu, 24 Jan 2019	Prob (F-statistic):	0.00			
Time:	13:53:58	Log-Likelihood:	-6.6118e+06			
No. Observations:	5800000	AIC:	1.322e+07			
Df Residuals:	5799996	BIC:	1.322e+07			
Df Model:	3					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	-0.2029	0.003	-58.657	0.000	-0.210	-0.196
gdpppp	0.8585	0.000	3193.736	0.000	0.858	0.859
gini_ref	-0.0177	4.04e-05	-437.275	0.000	-0.018	-0.018
c_i_parent	0.3213	0.000	945.316	0.000	0.321	0.322

coefficient de gini_ref (**-0.018**) :

si pour toute autres variables constante,

il augmente de **n** ,

income_child sera divisé par **$e^{n \times 0.018}$**

(exemple)

*Et si l'on ajoute la classe de **revenu des parents** ?*

Régression linéaire → **log(income_child)** à partir de **log(GDP PPP)**, **gini_ref** et **log(C_{i,parent})**

$$\text{SCT} = \text{SCE} + \text{SCR}$$

11 062 517 7 742 646 3 319 870

Selon le modèle,

$$R^2 = 0.70$$

70% de la **variation de revenu** est expliquée par
le **pays de naissance** et le **revenu des parents**

Conclusion

- Coefficients significatifs
- Homoscédasticité
- Variables non-colinéaires

$$R^2 = 0.70$$

Régression linéaire

revenu parents
revenu moyen de son pays
indice de Gini de son pays



revenu de l'enfant

Prédiction

classe 77
30 357 \$
32.1



21 852 \$